INF 558: Building Knowledge Graph

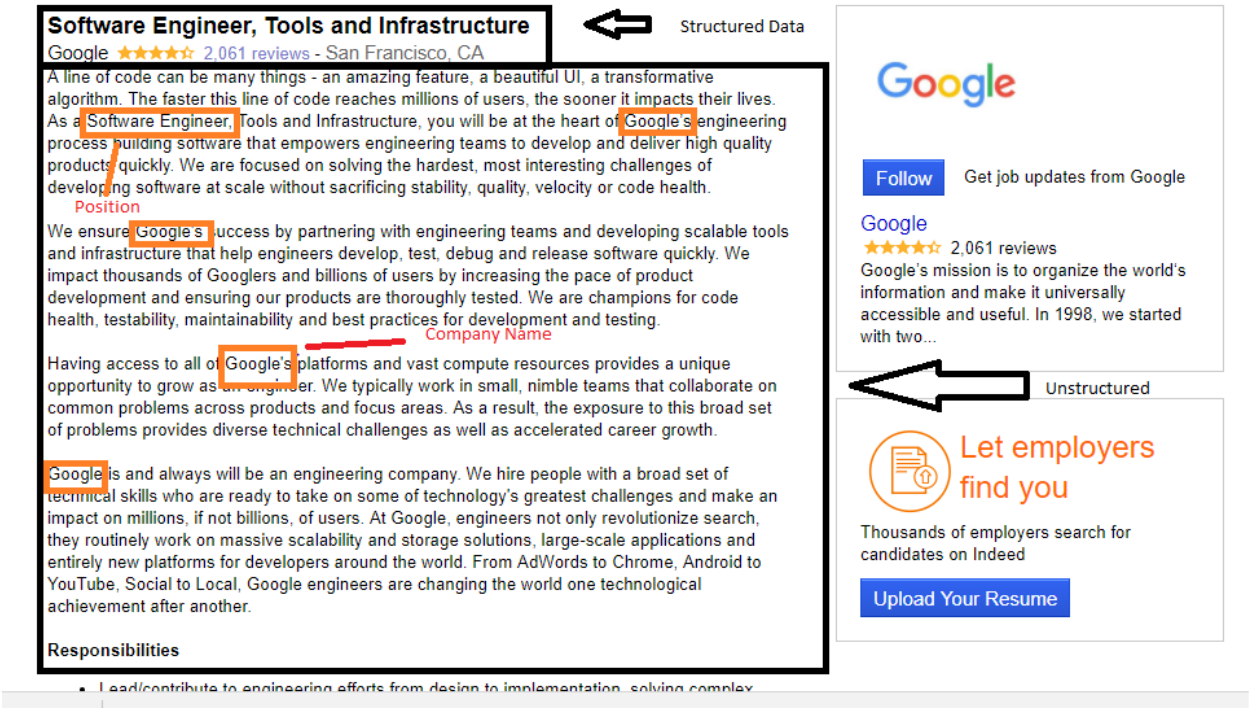
Homework 4: CRF

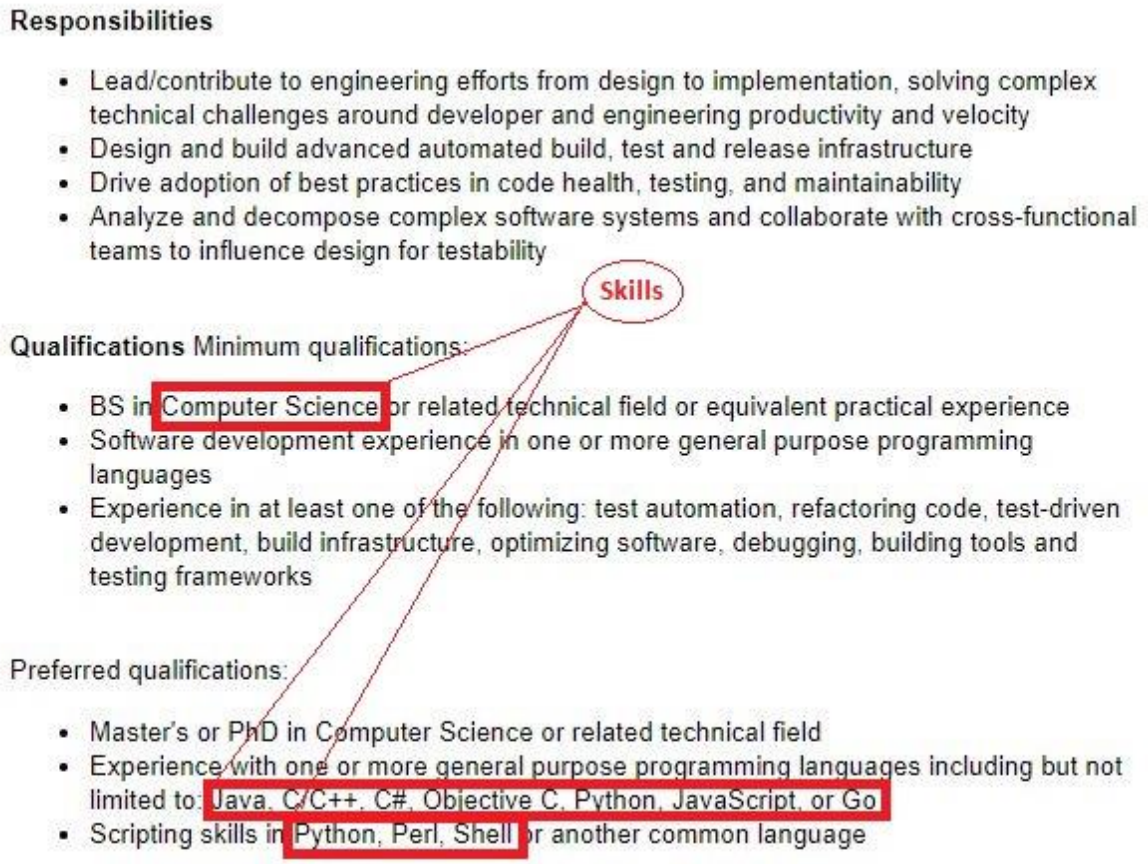
Q.1) What was the information you were looking for? Describe the labels you chose and why.Also include at least one screenshot of the webpage you are using and show where is the information you are looking for.

My source of unstructured data is from indeed.com, which is a job portal. Also my previous assignments retrieved structural data and our project is based on scraping from job portals, hence this selection.
I have used scrapy to crawl the job posting and retrieve the unstructured job description.
Scrapy spider: indeed.py
I have stored the output of this crawled data in a json file: Abhishek_Dhameja_hw4.json.

The following are the key attributes I planned to retrieve from the unstructured data:

1. Company Name (C)
2. Position (P)
3. Skills (S)





Q.2) What kind of tags did you tag your data with? Explain your choices.

Following tags are collected from the Job description which are scraped from the website:

1. Company Name (tagged as "C")
2. Position (tagged as "P")
3. Skill (tagged as "S")
4. Irrelevant data (tagged as "I") (the data which do not belong to any of the above three categories)

I have manually marked 50 training records and 20 test records with C/P/S tag.
Build the CRF model on the train data.
Predicted and compared with the manually tagged test data.

Q.3) Report your classifier's precision, recall and F-1 measure. Why did your classifier perform well (or not satisfactorily)?

Find the snapshot of the classifier's performance :

```
C:\Python27\python.exe E:/Workspaces/Python/inf558/Abhishek_Dhameja_hw4/Abhishek_Dhameja_crf.py
             precision    recall  f1-score   support

          C       1.00      0.09      0.17        65
          I       0.97      1.00      0.98      8361
          P       0.52      0.38      0.44        29
          S       0.84      0.28      0.42       304

avg / total       0.96      0.96      0.95      8759


Process finished with exit code 0
```

1. Classifier performs decently in case of the P and the S tag, as all the jobs did mention in the description the position and the skill.
2. It performs poorly in case of the C tag as most of the company have not mentioned their names in the description(as already present in the structured heading)(We can retrieve that information from the heading)
3. The classifier's performance is boosted as I have used POS tags as one of the features among others